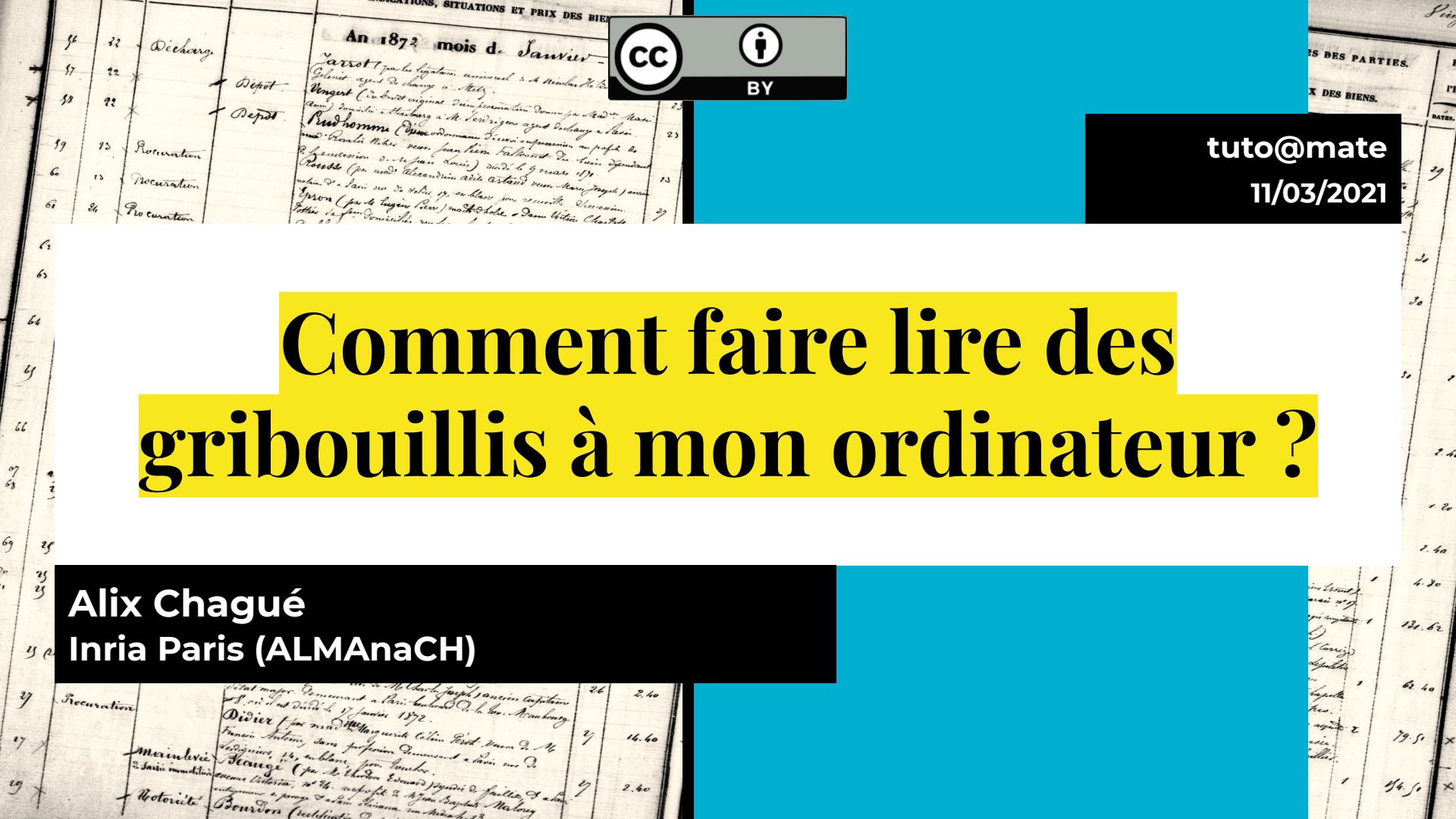
Comment faire lire des gribouillis à mon ordinateur ?

Alix Chagué



tuto@mate
11/03/2021

Comment faire lire des gribouillis à mon ordinateur ?

Alix Chagué
Inria Paris (ALMAnaCH)

ALMAnaCH contribue au projet eScriptorium (SCRIPTA-PSL)

Inria ALMAnaCH

projet DAHN

projet CREMMA

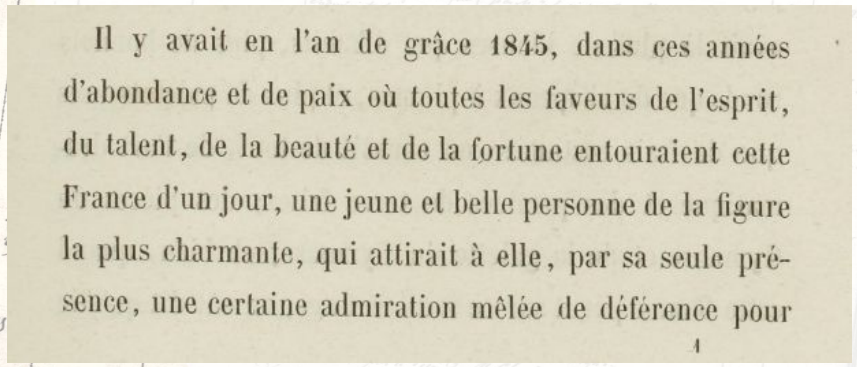
projet LECTAUREP

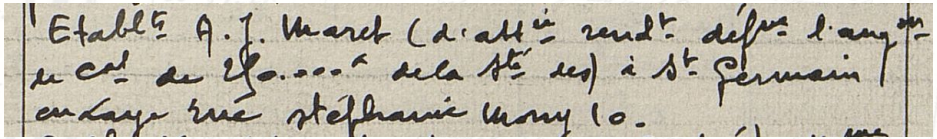
- Transcription automatique de documents d'archive
- Édition numérique
- Indexation automatique et fouille de données
- Mise en production d'une infrastructure publique

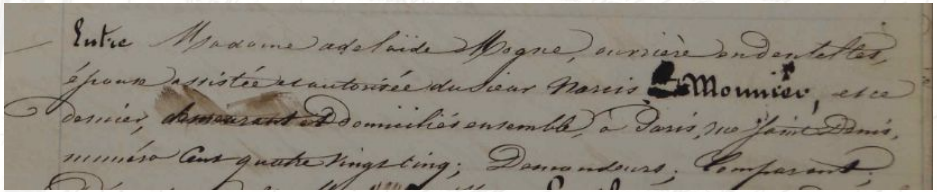
Définitions

Définitions : OCR vs HTR ?

- **OCR** : Optical Character Recognition ↠ reconnaissance du texte en général (imprimé principalement)
- **HTR** : Handwritten Text Recognition ↠ reconnaissance du texte manuscrit (qui prend en compte les difficultés inhérentes à ces textes)

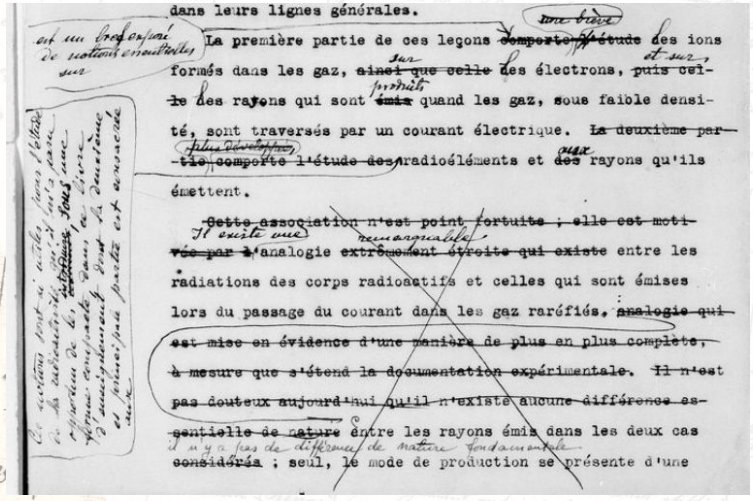

Il y avait en l'an de grâce 1845, dans ces années d'abondance et de paix où toutes les faveurs de l'esprit, du talent, de la beauté et de la fortune entouraient cette France d'un jour, une jeune et belle personne de la figure la plus charmante, qui attirait à elle, par sa seule pré-





Un logiciel « d'HTR » est capable de lire du texte imprimé !

Définitions : OCR vs HTR ?



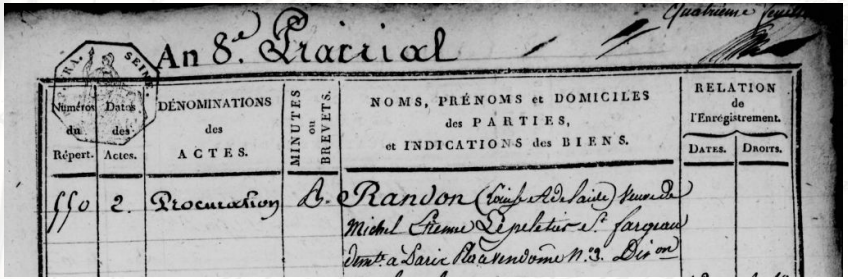
dans leurs lignes générales.

La première partie de ces leçons ~~comporte l'étude~~ des ions formés dans les gaz, ~~ainsi que celle~~ des électrons, ~~puis celle~~ des rayons qui sont ~~émis~~ quand les gaz, sous faible densité, sont traversés par un courant électrique. ~~La deuxième partie comporte l'étude des~~ radioéléments et ~~des~~ rayons qu'ils émettent.

~~Cette association n'est point fortuite ; elle est motivée par l'~~analogie ~~extrêmement étroite qui existe~~ entre les radiations des corps radioactifs et celles qui sont émises lors du passage du courant dans les gaz raréfiés. ~~Analogie qui est mise en évidence d'une manière de plus en plus complète, à mesure que s'étend la documentation expérimentale. Il n'est pas douteux aujourd'hui qu'il n'existe aucune différence essentielle de nature~~ entre les rayons émis dans les deux cas ~~considérés~~ ; seul, le mode de production se présente d'une

- Différentes orientations du texte
- Différents sens de lecture du texte (GàD, DàG, HeB, BeH)
- Différents alphabets

- Grandes catégories d'écriture :
  - Imprimée
  - “Tapuscrite”
  - Manuscrite
- Elles peuvent être mélangées au sein d'un même document (épreuve corrigée à la main, formulaire pré-imprimé, ...)



An 8e Prairial

Numéros du Répert.	Dates des Actes.	DÉNOMINATIONS des ACTES.	MINUTES ou BREVETS.	NOMS, PRÉNOMS et DOMICILES des PARTIES, et INDICATIONS des BIENS.	RELATION de l'Enregistrement. DATES.	DROITS.
550	2.	Procuration	A.	Randon (Louise Adelaïde Sainte) veuve de Michel Etienne Répétiteur [illegible] de Sarguemines Demte. à Paris Rue Vendôme N.° 3. Div. en		

De la REM, pour quoi faire ?

L'objectif de la **reconnaissance automatique d'écriture manuscrite (REM)** est principalement :

- d'accélérer la transcription ;
- Donc de la rendre moins pénible ;
- Donc de traiter de plus grands ensembles documentaires.

Mais la REM n'est pas une fin en soi :

- Édition numérique ?
- Analyse stylistique ou fouille de texte ?
- Accessibilité des ressources (moteur de recherche) ?
- etc.

Quelques logiciels d'HTR

Transkribus

K4all + Calamari OCR

eS + KRAKEN

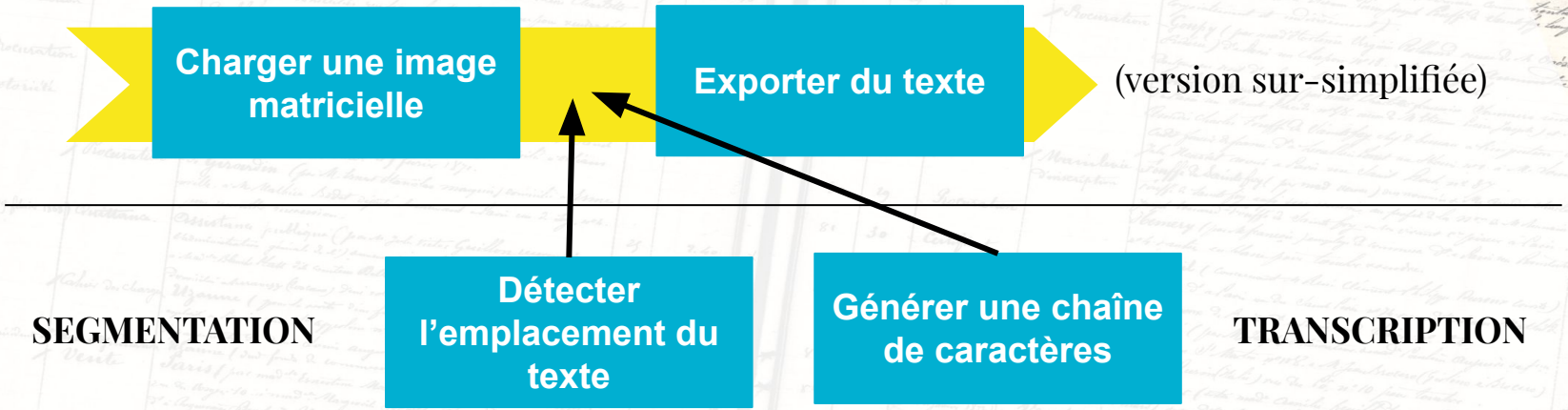
TESSERACT

Voir le recensement réalisé par Awesome-OCR :
https://github.com/kba/awesome-ocr

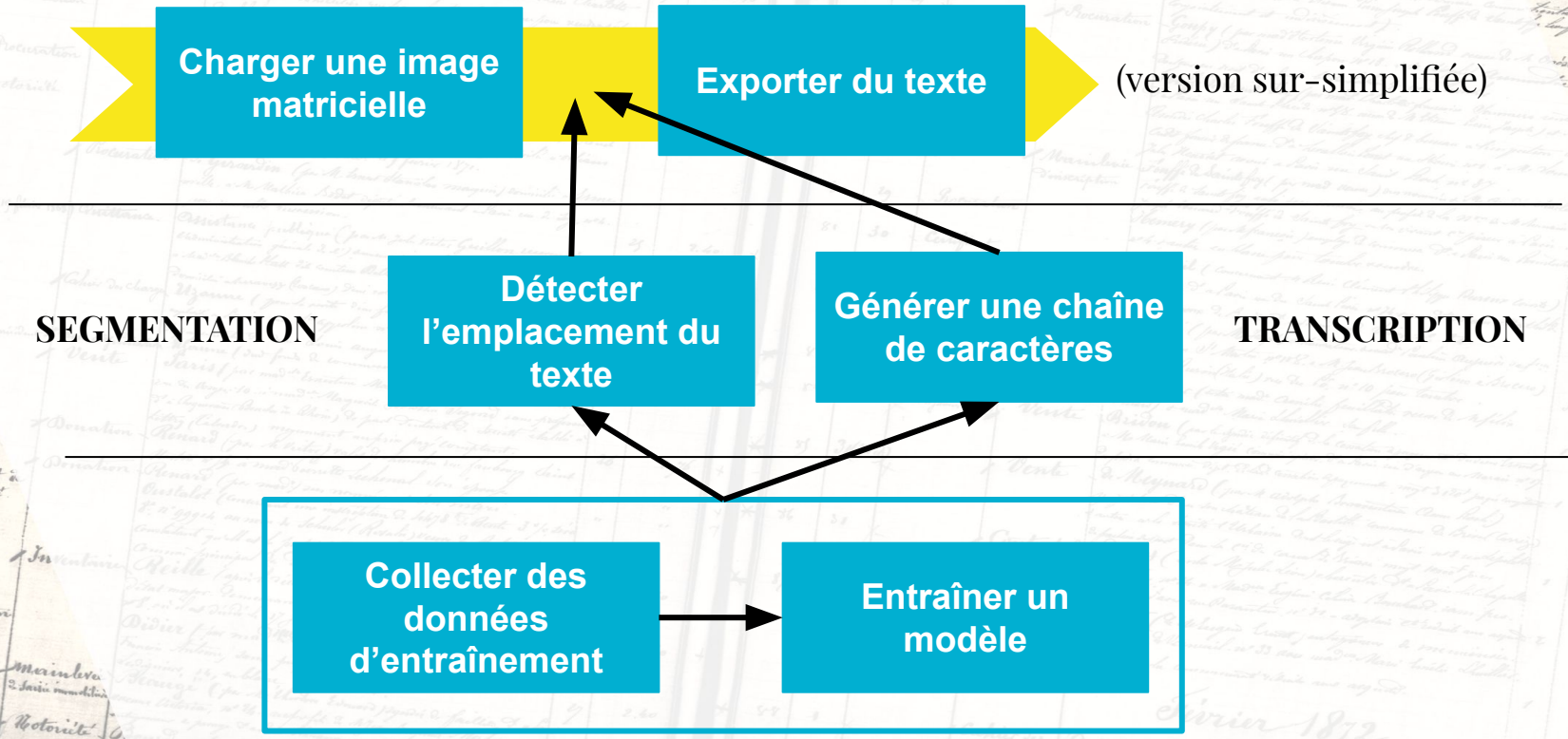
Schématisation du workflow

(version sur-simplifiée)

Schématisation du workflow



Schématisation du workflow



Deux workflows à ne pas confondre

La transcription doit être la plus parfaite possible mais pas nécessairement exhaustive

A On part de zéro :

- Il faut créer (ou collecter) des données d'entraînement (***vérité terrain***)
- Puis entraîner un ou des modèles de segmentation et/ou de transcription)

B On a déjà un modèle de transcription / segmentation :

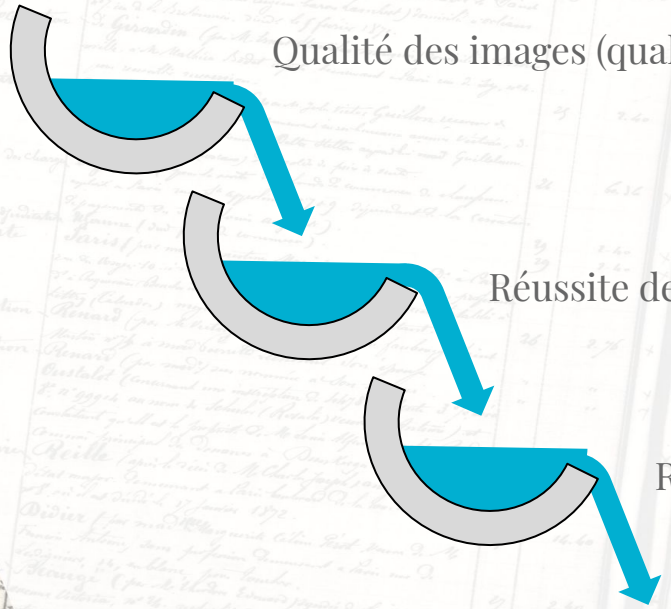
- Simplement appliquer le(s) modèle(s) aux images et vérifier le résultat

On vise l'exhaustivité mais il ne faut pas nécessairement attendre une transcription parfaite

Si le modèle n'est pas suffisamment bon, on bascule dans le premier cas de figure

Attention à ne pas négliger l'étape de la segmentation

La qualité de la segmentation est aussi importante que la transcription.



Qualité des images (qualité des documents, bruits, etc)

Réussite de la segmentation (faux positifs, faux négatifs,...)

Réussite de la transcription

Un gain de temps ?

La REM accélère la tâche de transcription mais pas celles :

- ❖ De la vérification des images
- ❖ De la vérification de la segmentation
- ❖ De la vérification de la transcription

=> elles ne sont pas prises en charge dans le cadre de la REM pure.

Ces tâches peuvent être confiées à des personnes aidées d'outils de pré/post-traitement.

Exemples :

- ScanTailor pour le découpage des images
- PySpellchecker pour le post-traitement, etc

Plusieurs paramètres sont à considérer :

- ❖ La complexité des écritures et des documents ;
- ❖ Les objectifs d'exploitation de cette transcription ;
- ❖ L'existence d'autres corpus similaires déjà annotés ou de modèles correspondants au corpus.

Il faut donc préparer la campagne d'HTR en amont en prenant en compte ces éléments !

Préparer une campagne de REM

Pourquoi préparer une campagne

- Pour choisir le logiciel le plus adapté ;
- Pour connaître et anticiper les points de blocages ;
- Pour adapter la stratégie d'annotation aux difficultés et aux objectifs (et éviter d'inutiles retours en arrière) ;
- Pour pouvoir créer des jalons pour évaluer l'avancée de la campagne ;

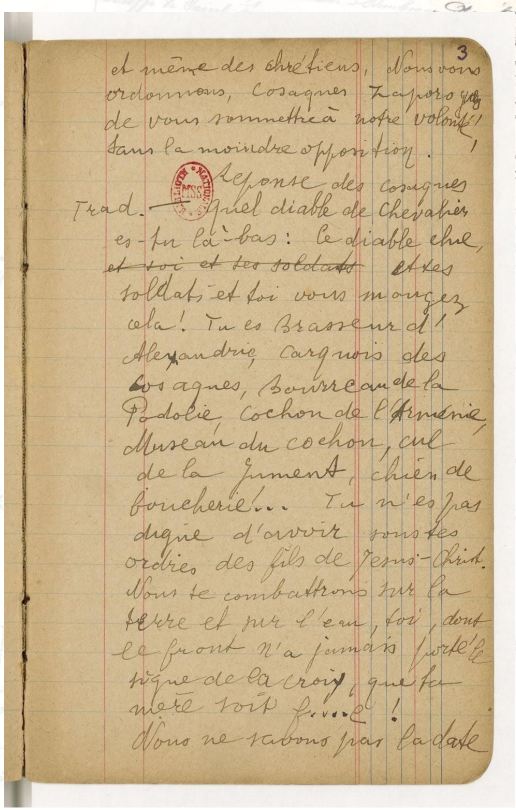
Anticiper les difficultés : évaluer le corpus

- Quel type d'écriture ? Y a-t-il plusieurs écritures ? Combien ? Varient-elles beaucoup ?
- Quelle est la qualité des numérisations (résolution, cadrage, binarisé…) ?
- Quelle est la qualité des documents (déchirures, taches, plis, etc) ?
- Y a-t-il des abréviations dans le texte ? Y a-t-il des symboles particulier ?
- Y a-t-il autre chose que du texte (des illustrations, des dessins, des tableaux) ? Comment cela s'articule-t-il avec le texte ?
- Y a-t-il des décoration sur le texte (surlignage, soulignage, etc) dont je veux tenir compte ?
- Etc.

Anticiper les difficultés : exemple

J'ai 1 000 pages d'un carnet personnel du XXe siècle et je veux pouvoir faire de la fouille de texte :

- ★ Des numérisations couleur en haute résolution
- ★ Une dizaine de pages gondolées et abîmées par l'eau
- ★ Une seule main d'écriture
- ★ Des abréviations et des mots mal formés
- ★ Quelques symboles "maisons"
- ★ Quelques dessins dans les marges
- ★ Des mots importants soulignés, des mots barrés



Anticiper les difficultés : adapter les objectifs

En fonction des objectifs du projet de transcription, trier les critères :

- Ce qui dépend de la REM et ce qui n'en relève pas
- Ce qu'il est fondamentale de réussir à détecter (ce qui n'a pas besoin d'être segmenté ou transcrit)
- Quel degré de perfection est nécessaire

Si on reprend l'exemple :

- ★ Seul modèle
- ★ Des caractères pour transcrire les symboles maisons
- ★ Les mots soulignés ou barrés ne seront pas traités différemment
- ★ Dessins ignorés
- ★ Prévoir des difficultés à lire les mots mal formés (corriger avec un script ou à la main après)
- ★ Pages gondolées traitées à la main

Définir des règles d'annotation

Seul-e ou en équipe : définir des règles d'annotation en amont !

- Expliciter les règles des cas "normaux" ;
- Définir des règles précises et univoques pour les cas particuliers ;
- Définir des comportements pour les cas "anormaux" imprévus ;

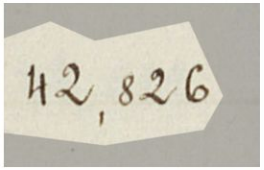
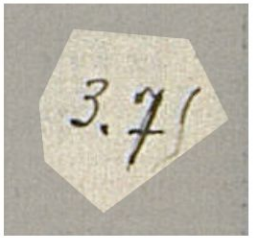
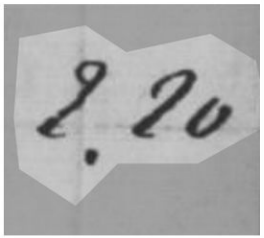
Pourquoi définir des règles ?

- Faciliter le contrôle qualité ;
- Assurer la constance de l'annotation dans le temps et entre les annotateurs ;
- Documenter l'annotation pour de futurs utilisateurs ;

Définir des règles d'annotation : exemples

- Comment annoter les mots barrés ou illisibles ?
- Comment traiter les abréviations ? Les fautes ?
- Comment traiter les caractères spéciaux ?
- Quel tiret utiliser pour transcrire un tiret ?
- Comment distinguer manuscrit et imprimé ?

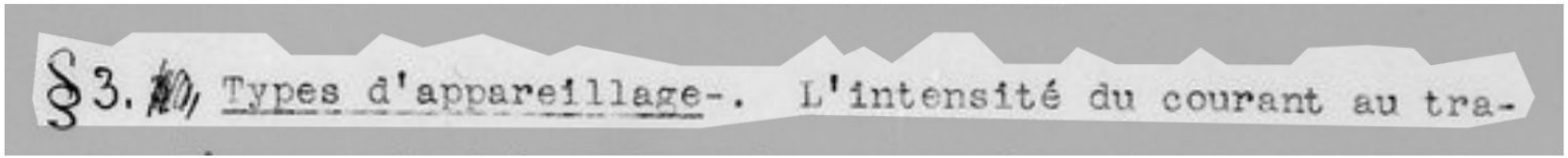
Transcription des nombres

Exemple	Transcription	Explication
	42,826	Le scribe a employé **une virgule** pour séparer la partie entière et la partie fractionnaire.
	3.75	Le scribe a employé **un point** pour séparer la partie entière et la partie fractionnaire.
	2.20	Le scribe a employé **un point** pour séparer la partie entière et la partie fractionnaire, bien que l'alignement avec la base du tracé des chiffres n'est pas parfait.

Annoter la mise en forme : exemple

- Les mots soulignés sont précédés d'un "_"
- Les mots barrés ne sont pas distingués des autres
- Les mots illisibles sont remplacés par un "#"
- Les mots écrits à la machine sont distingués de ceux écrits à la main car ils sont systématiquement précédés d'un symbole : "·"

A l'aide d'un script, il serait facile de transformer ces indicateurs en annotations xml en phase de post-traitement.



§3 # _·Type _·d'appareillage-. ·L'intensité ·du ·courant ·au ·tra-

Remarques sur la gestion d'un projet de REM

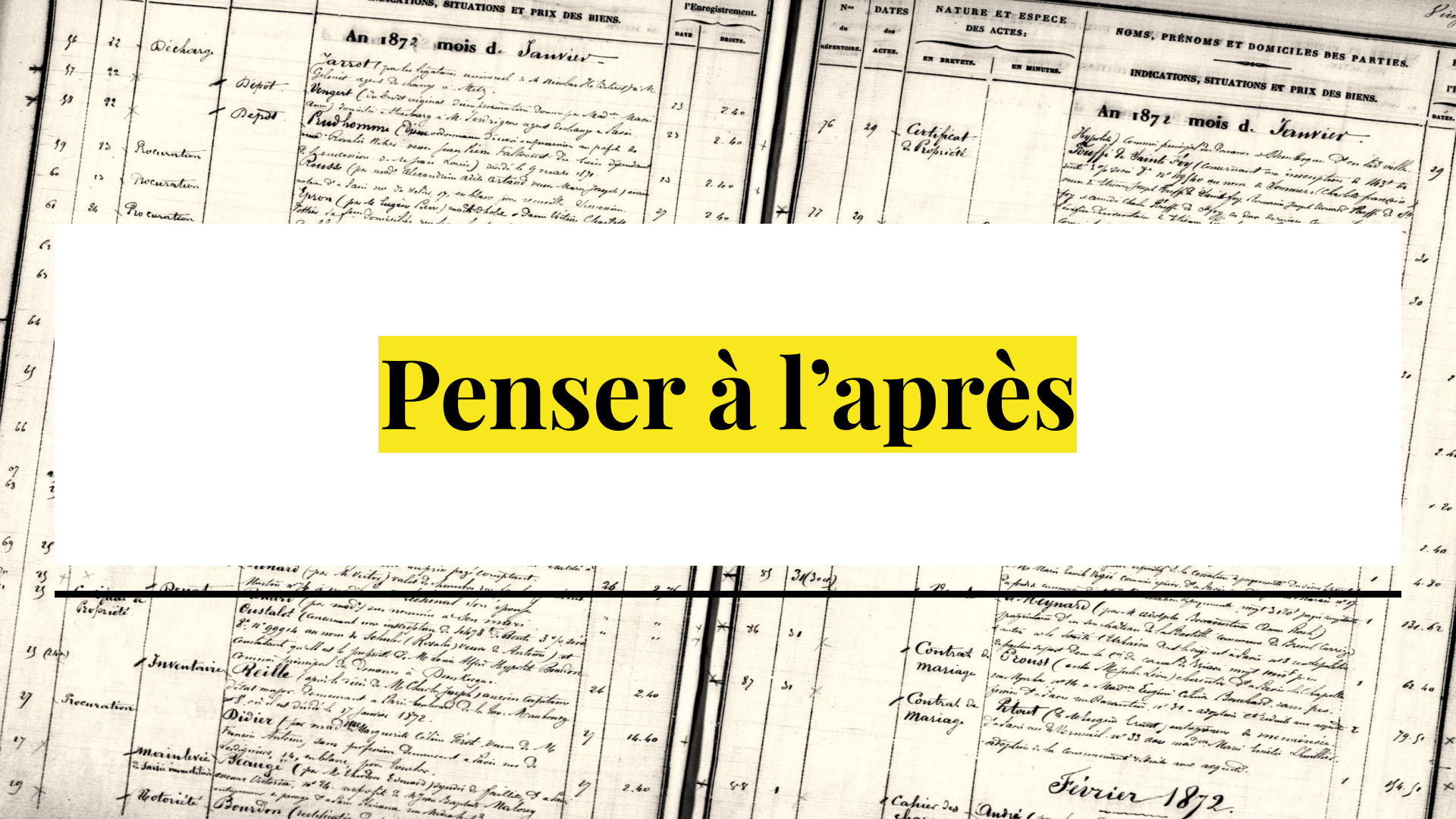
Objectif général : obtenir la transcription de la totalité du corpus

Mais :

- ➢ Avec quel niveau de perfection ?
- ➢ Avec quels jalons ?

Quelques exemples de jalons dans le scénario B :

- ❖ Créer un set d'évaluation avancée représentatif du corpus et des difficultés identifiées (↞ sc. A & B)
- ❖ Annoter et controler *x* pages/lignes de vérité terrain (× nombre de modèles nécessaires) (↞ sc. B)
- ❖ Entraîner un/des modèle(s) et tester la performance effective sur le set d'évaluation avancée (↞ sc. A & B)
- ❖ Développer/sélectionner les outils de post-traitement (correction/nettoyage/augmentation) (↞ sc. A & B)



Penser à l'après

Des données intermédiaires à ne pas jeter

La vérité de terrain est un produit dérivé précieux et qui fait partie des données de la recherche !

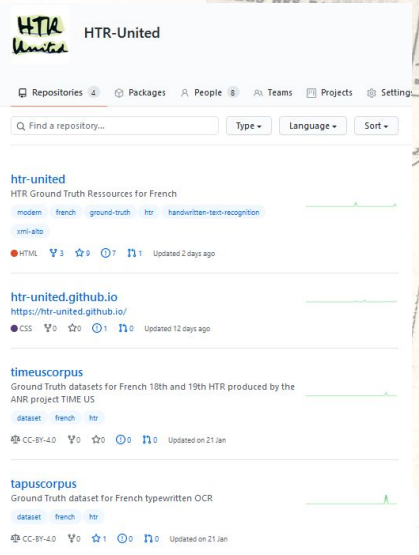
- Elle peut être utile à d'autres projets disposant d'écritures similaires
- Plus on a de données disponibles, moins on perd de temps pour l'entraînement d'un modèle !

Un package minimal pour le partage :

- Images (ou liens vers les images)
- Transcription dans un format adapté (XML ALTO ou XML PAGE)
- Documentation sur les règles d'annotations et les écritures

HTR-United : une solution de partage des données

- Une organisation Github pour partager vos données sous la forme de repository si vous ne savez pas où les publier
- Un point de centralisation des signalement de sets de données (un fichier de référencement dans le projet principal)




```yaml
name: htr-united
description: 'HTR Ground Truth Resources'
entries:
    -
        title : 'tapuscorpus'
        url: 'https://github.com/HTR-United/tapuscorpus'
        description: 'Ground Truth dataset for French typewritten OCR (20th century documents)'
        language: French
        time: 1900--1999
        hands: 30
        license:
            - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
        format: 'XML-Page'
        volume:
            - {count: "150", metric: pages}
```

https://htr-united.github.io/


Démo

Liens vers de la documentation

Tutoriel pas à pas :

https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium

Démo vidéo :

https://escripta.hypotheses.org/escriptorium-video-gallery

Questions d'infrastructure

Le projet CREMMA

Mutualiser l'investissement infrastructurel pour rendre une solution open-source comme eScriptorium plus accessible

Inria
ALMAnaCH

\+ ENC ; EFEO ; EPHE ; IRHT ; LAMOP

Financement DIM MAP (IDF)

eScriptorium
SCRIPTA
PSL

- RAM
- GPU
- Stockage
- Maintenance

Déploiement sur Serveur Inria

Merci !